USGS
science for a changing world

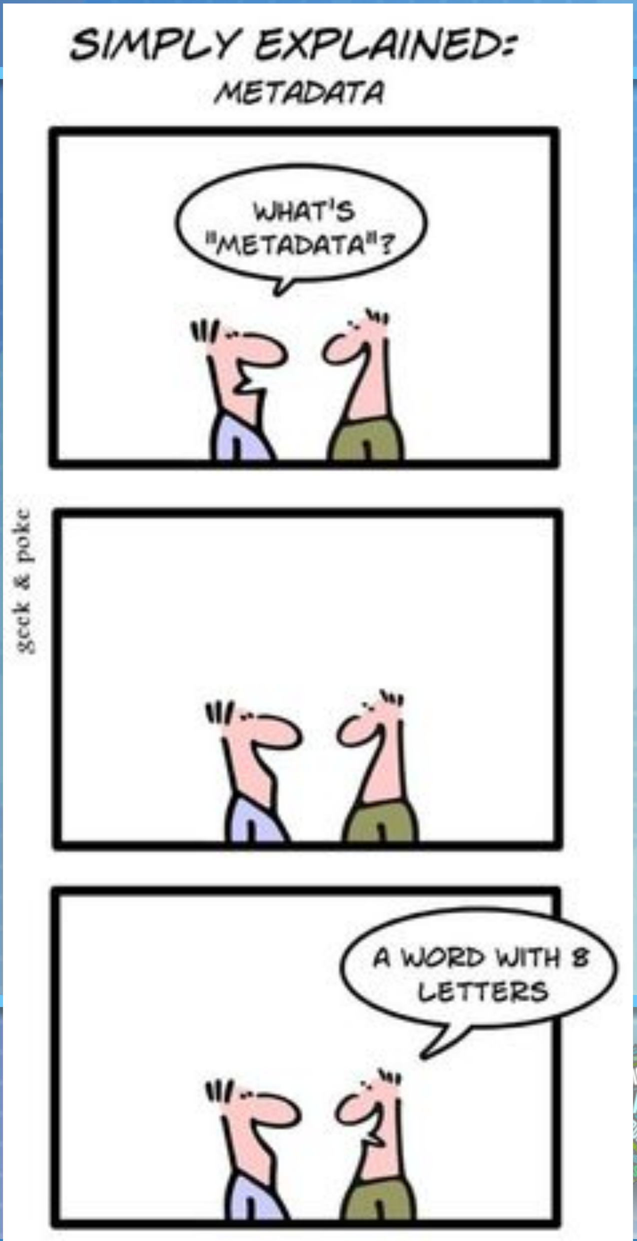
Metadata: standards, tools, and best practices

Lisa Zolly
Core Science Analytics, Synthesis & Libraries
PNAMP & USGS CDI Data Management Series
15 July 2015

U.S. Department of the Interior
U.S. Geological Survey

The five stages of metadata

Denial



Geek & Poke http://geek-and-poke.com. Used under CC License.

The five stages of metadata

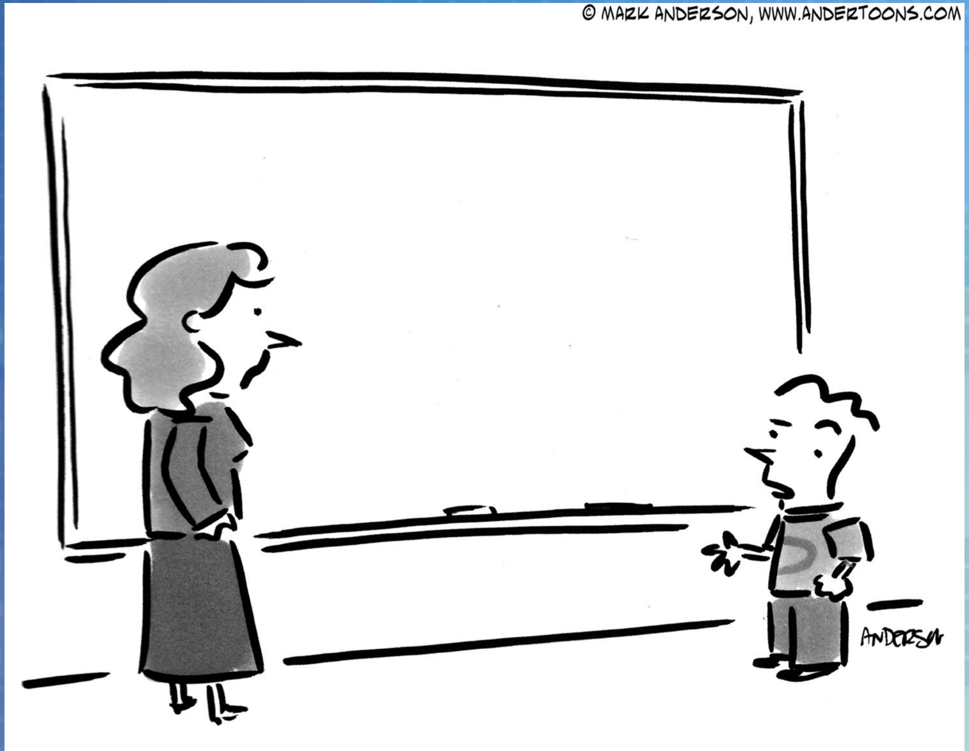
Anger

"You want my metadata? You'll have to pry it from my cold, dead hands..."
TimoElliott.com

Image: Timo Elliott, CC License. http://timoelliott.com Image used with permission.

The five stages of metadata

Bargaining



"Before I write my name on the board, I'll need to know how you're planning to use that data."

The five stages of metadata

Depression

"I liked it better before big data and metadata when we just had good old regular data."

Image © Mark Anderson, Andertoons.com. Licensed for this presentation. Do not copy.

USGS

The five stages of metadata

Acceptance

Yes, this may be the world's nerdiest t-shirt.

(Yes, it's my t-shirt.)

T-shirt: Café Press, riffing Metallica's logo. (Note to Metallica's lawyer: I didn't design it.)

Metadata helps to make data

✓Discoverable

✓Understandable

✓Defensible

Metadata for discovery (metadata catalogs)

- ✓ Who
- ✓ What
- ✓ Where
- ✓ When

DATA.GOV

ORNL DAAC
DISTRIBUTED ACTIVE ARCHIVE CENTER
FOR BIOGEOCHEMICAL DYNAMICS

data.noaa.gov

GEOPLATFORM.gov

NASA Distributed Active Archive Center (DAAC) at NSIDC

ONEMercury A DataONE Search Tool for Scientific Data

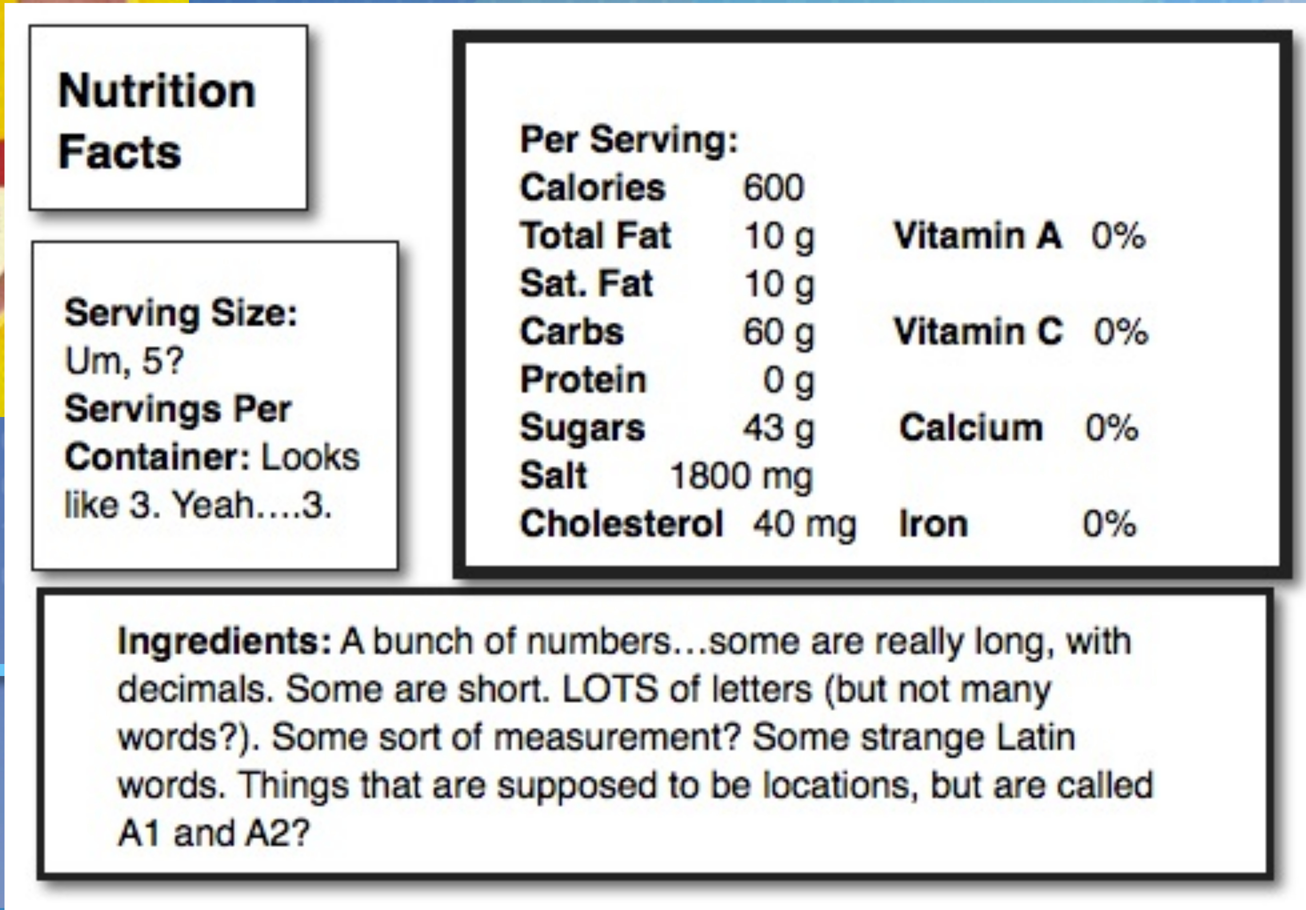
U.S. Geological Survey Science Data Catalog

NATIONAL
FISH HABITAT
PARTNERSHIP
NFHP Data System Home >> Map Viewers

ARM
CLIMATE RESEARCH FACILITY

Metadata for understanding

Good metadata prevents your data from becoming 'mystery meat'



Nutrition Facts

Serving Size: Um, 5?
Servings Per Container: Looks like 3. Yeah....3.

Per Serving:

Calories	600		
Total Fat	10 g	**Vitamin A**	0%
Sat. Fat	10 g		
Carbs	60 g	**Vitamin C**	0%
Protein	0 g		
Sugars	43 g	**Calcium**	0%
Salt	1800 mg		
Cholesterol	40 mg	**Iron**	0%

Ingredients: A bunch of numbers...some are really long, with decimals. Some are short. LOTS of letters (but not many words?). Some sort of measurement? Some strange Latin words. Things that are supposed to be locations, but are called A1 and A2?

Metadata for understanding

Data ‘lake’: Beautiful, well described, well managed, useful, useable data

Data ‘swamp’:*
Data that are unusable ‘clutter’
with no further value

*With apologies to wetlands biologists for the metaphor

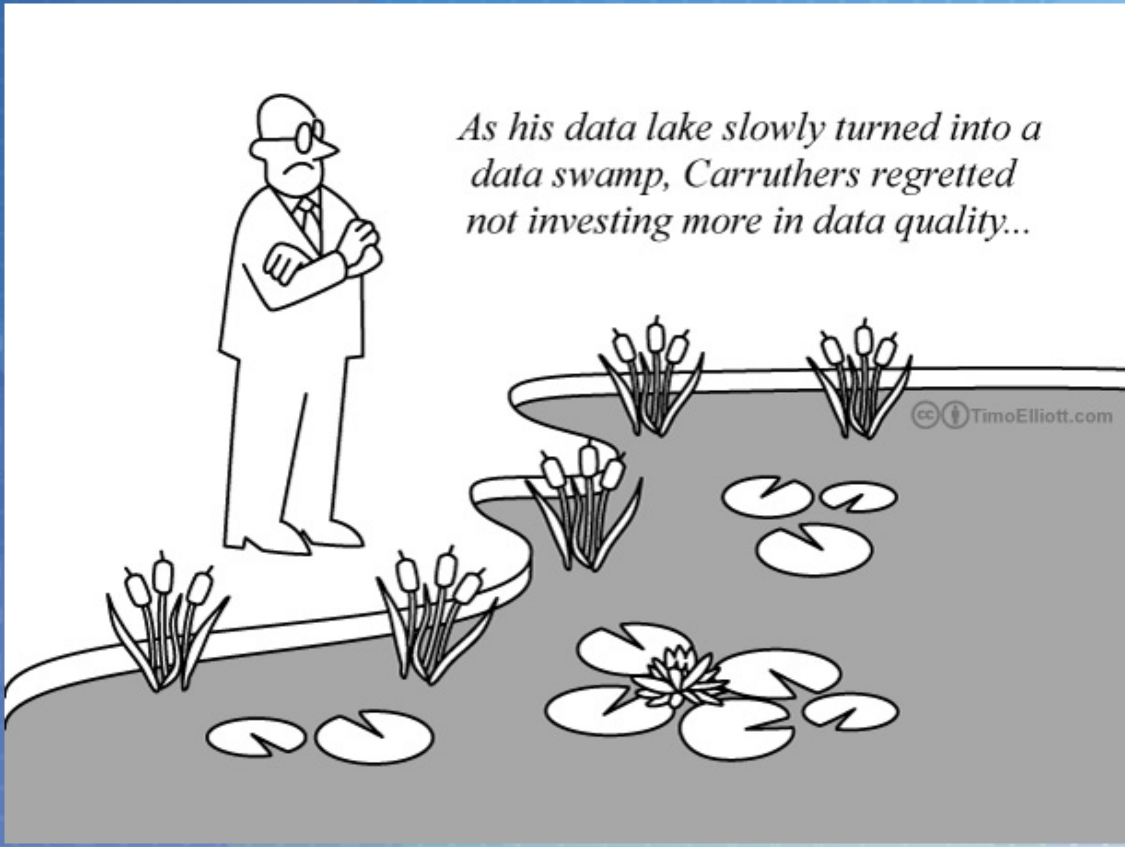
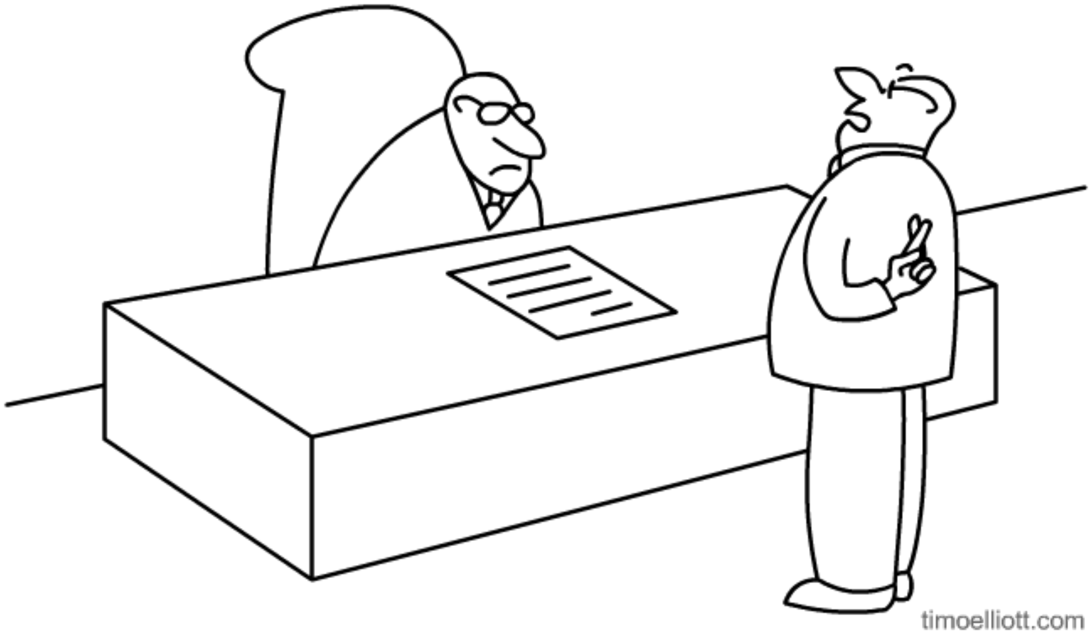

As his data lake slowly turned into a data swamp, Carruthers regretted not investing more in data quality...
TimoElliott.com

Image: Timo Elliott, CC License. http://timoelliott.com Image used with permission.

Metadata for (defensible) reuse

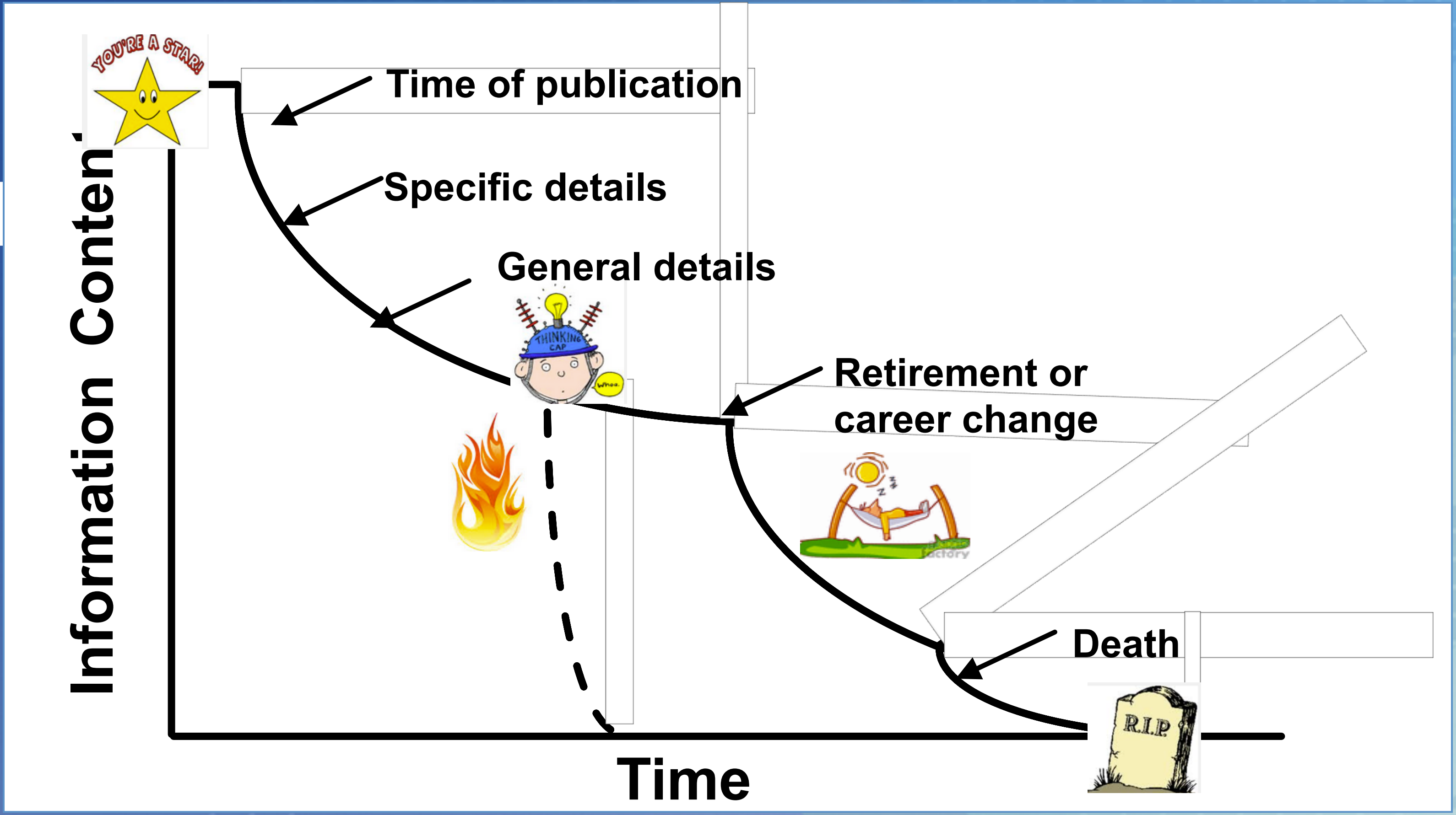
Can your metadata help to support responsible reuse of your data?



"Yes sir, you can absolutely trust those numbers"

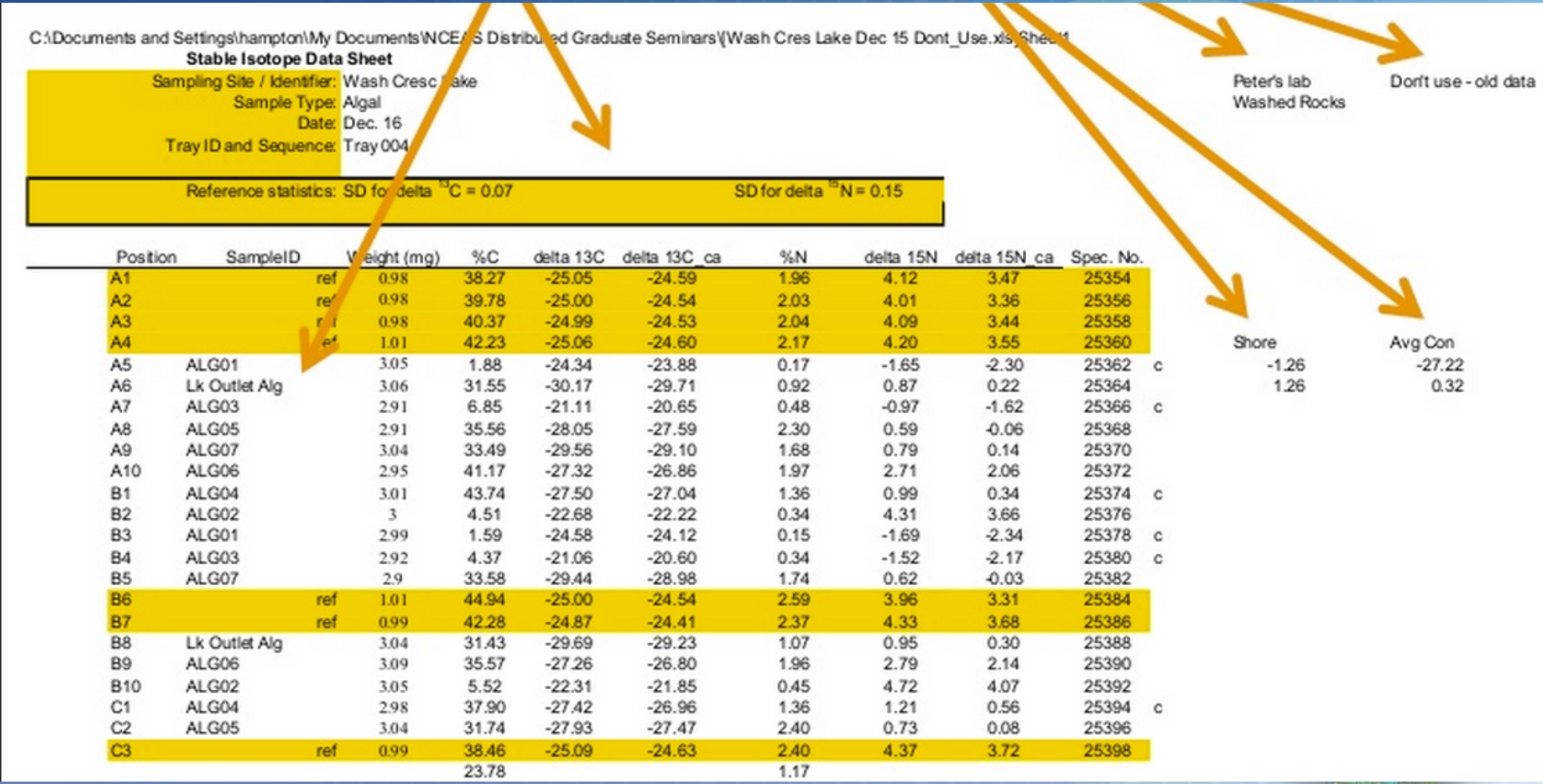
Image: Timo Elliott, CC License. http://timoelliott.com Image used with permission.

What's the useful life of your data in the absence of good metadata?



Graphic: Michener, et al., 1997

Can users understand your data? Will *you* understand your data in 5 years? In 10?



C:\Documents and Settings\hampton\My Documents\NCEAS Distributed Graduate Seminars\[Wash Cres Lake Dec 15 Dont_Use.xls]Sheet1

Stable Isotope Data Sheet

Sampling Site / Identifier:	Wash Cresc Lake
Sample Type:	Algal
Date:	Dec. 16
Tray ID and Sequence:	Tray 004

Reference statistics: SD for delta ^{13}C = 0.07 SD for delta ^{15}N = 0.15

Peter's lab
Washed Rocks

Don't use - old data

Position	SampleID	Weight (mg)	%C	delta 13C	delta 13C_ca	%N	delta 15N	delta 15N_ca	Spec. No.			
A1	ref	0.98	38.27	-25.05	-24.59	1.96	4.12	3.47	25354			
A2	ref	0.98	39.78	-25.00	-24.54	2.03	4.01	3.36	25356			
A3	ref	0.98	40.37	-24.99	-24.53	2.04	4.09	3.44	25358			
A4	ref	1.01	42.23	-25.06	-24.60	2.17	4.20	3.55	25360		Shore	Avg Con
A5	ALG01	3.05	1.88	-24.34	-23.88	0.17	-1.65	-2.30	25362	c	-1.26	-27.22
A6	Lk Outlet Alg	3.06	31.55	-30.17	-29.71	0.92	0.87	0.22	25364		1.26	0.32
A7	ALG03	2.91	6.85	-21.11	-20.65	0.48	-0.97	-1.62	25366	c		
A8	ALG05	2.91	35.56	-28.05	-27.59	2.30	0.59	-0.06	25368			
A9	ALG07	3.04	33.49	-29.56	-29.10	1.68	0.79	0.14	25370			
A10	ALG06	2.95	41.17	-27.32	-26.86	1.97	2.71	2.06	25372			
B1	ALG04	3.01	43.74	-27.50	-27.04	1.36	0.99	0.34	25374	c		
B2	ALG02	3	4.51	-22.68	-22.22	0.34	4.31	3.66	25376			
B3	ALG01	2.99	1.59	-24.58	-24.12	0.15	-1.69	-2.34	25378	c		
B4	ALG03	2.92	4.37	-21.06	-20.60	0.34	-1.52	-2.17	25380	c		
B5	ALG07	2.9	33.58	-29.44	-28.98	1.74	0.62	-0.03	25382			
B6	ref	1.01	44.94	-25.00	-24.54	2.59	3.96	3.31	25384			
B7	ref	0.99	42.28	-24.87	-24.41	2.37	4.33	3.68	25386			
B8	Lk Outlet Alg	3.04	31.43	-29.69	-29.23	1.07	0.95	0.30	25388			
B9	ALG06	3.09	35.57	-27.26	-26.80	1.96	2.79	2.14	25390			
B10	ALG02	3.05	5.52	-22.31	-21.85	0.45	4.72	4.07	25392			
C1	ALG04	2.98	37.90	-27.42	-26.96	1.36	1.21	0.56	25394	c		
C2	ALG05	3.04	31.74	-27.93	-27.47	2.40	0.73	0.08	25396			
C3	ref	0.99	38.46	-25.09	-24.63	2.40	4.37	3.72	25398			
			23.78			1.17						

Example file courtesy Stephanie Hampton

Why do we need to follow a metadata standard? (or, what's wrong with README files?)

- ✓ Ensures your data are fully described to facilitate understanding and reuse
  - ✓ People who understand data developed these standards
- ✓ Structures your metadata as machine readable XML
- ✓ Ingest of standardized metadata in catalogs helps enable persistent discoverability

Metadata standards for scientific data include

- ✓ ISO 19115 ‘suite’*
- ✓ Content Standard for Digital Geospatial Metadata (CSDGM), a.k.a. ‘FGDC’*
- ✓ Ecological Metadata Language (EML)
- ✓ NetCDF

* Official FGDC-endorsed standards for federally funded, geospatial data

USGS

A bit about EML and netCDF, since I won't be addressing them specifically

EML

- ✓ Used primarily by academia, NGOs
- ✓ Tools include Morpho and MERMAid

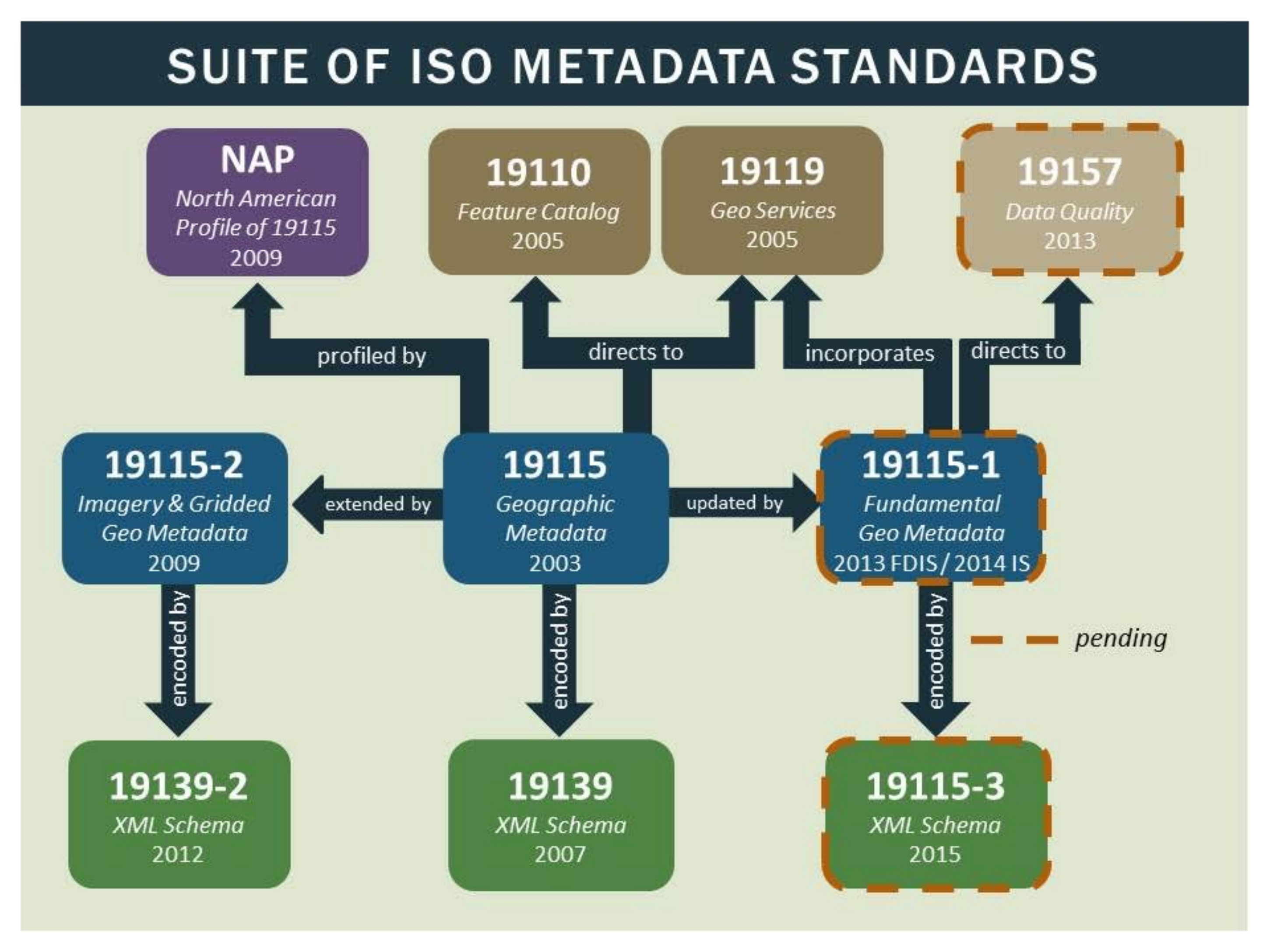
netCDF

- ✓ Supports climate data
- ✓ UCAR and NSIDC provide listings of compatible tools
- ✓ NOAA supports tools that generate ISO 19115-2 records from netCDF files

ISO metadata

- ✓ International suite of standards
- ✓ Endorsed by FGDC
- ✓ Small pockets of implementation in the federal sector to date
- ✓ Emphasis on machine-readable metadata files
- ✓ Reusable modules for repeated elements (e.g., contacts)

‘ISO metadata’ is a modular suite of standards



ISO metadata

- ✓ More ‘flexible’ due to fewer mandatory elements
- ✓ Does a better job of describing services, models
- ✓ Facilitates effective documentation of relationship between datasets and collections (child<->parent)

ISO tools

- ✓ Native ISO tools are lacking
- ✓ No single software solution to implement full suite of the standard
- ✓ Most available tools focus on the 19115.x module:
  - ✓ GRIIDC ISO 19115-2 Editor
  - ✓ MERMAid can extract the 19115 equivalents from CSDGM and produce a 19139-2 record (which is the XML expression of a 19115-2 compliant record). MERMAid can also generate one section from the 19110 module.
  - ✓ ArcGIS for Desktop can generate ‘ESRI ISO’ and a 19139-2 XML version of it.

ISO tools

- ✓ XML editors are the primary creation mechanism for full ISO metadata
  - ✓ Presumes high level of expertise with the standard
  - ✓ Requires comfort level generating native XML
- ✓ Validation tools also require expertise
- ✓ Challenge: producing a record that includes 19110 and 19157 elements
  - ✓ How can your data be understood without these details?

USGS

ISO metadata

"…federal agencies are encouraged to transition to ISO metadata as their agencies are able to do so."

FGDC website, 7/15/2015

Why the slow uptake for ISO?*

1. ‘Better the devil you know’ (CSDGM) defense
2. Complex standard with challenging syntax
3. Lack of form-centric tools
4. Current difficulties creating record of equal robustness to CSDGM

* ***You’ve entered The Lisa Opinion Zone***

Why the slow uptake for ISO?*

5. Lack of clarity as to the extent to which CSDGM Profiles and Extensions are addressed in ISO

✓ BDP users may want to check out NOAA's Workbook for 19115 – Biological Extensions

6. CSDGM→ISO conversion tools assume that the CSDGM XML *strictly* adheres to the CSDGM standard

7. Distraction of the Open Data requirements and deadlines (more readily addressed via #1)

8. Organizational structure ← → relationship to how metadata are generated and managed

* ***Yep, still in The Lisa Opinion Zone***

USGS

What factors might support a successful migration to ISO *today* by an organization?*

Organizations with more homogeneous data

- Homogeneity in data types, formats, science domain

Organizations with more centralized metadata support

✓ Concentrated pool of metadata specialists

✓ Much higher ROI in training, tools, processes

** You guessed it, still in The Lisa Opinion Zone*

USGS

CSDGM standard

It’s not dead yet!

> “Most NSDI stakeholders have long utilized the [CSDGM], which will continue to have a legacy for many years….It is recognized that the transition to ISO metadata will be occurring over the next few years.”
>
> *- FGDC website, 7/15/2015*

CSDGM standard

- ✓ Developed by the FGDC in the mid 1990s
- ✓ Still the predominant standard in use by most federal agencies
- ✓ Designed for raster, point, and vector data
- ✓ Includes a number of official profiles and extensions (Biological, Shoreline, Remote Sensing)

CSDGM tools

Currently there is a wider array of CSDGM tools to facilitate development of a *robust and complete* metadata record.

Metadata that support understanding and reuse of data are a critical component of the federal mandate for open data.

If you are more likely *today* to be able to meet the open data imperative via CSDGM, you probably should use that standard.*

*Back in the Lisa Opinion Zone

USGS

CSDGM tools

FREE:

- ✓ EPA Metadata Editor (plug-in for ESRI) %
- ✓ Metavist (desktop) #
- ✓ MERMAid (online) ^
- ✓ Metadata Wizard (plug-in for ESRI) %
- ✓ Online Metadata Editor (online)# %
- ✓ TKME (desktop)^ %

# Supports Biological Data Profile
^ Supports Biological Data Profile, Shoreline Profile, and Remote Sensing Extension
% Provides validation

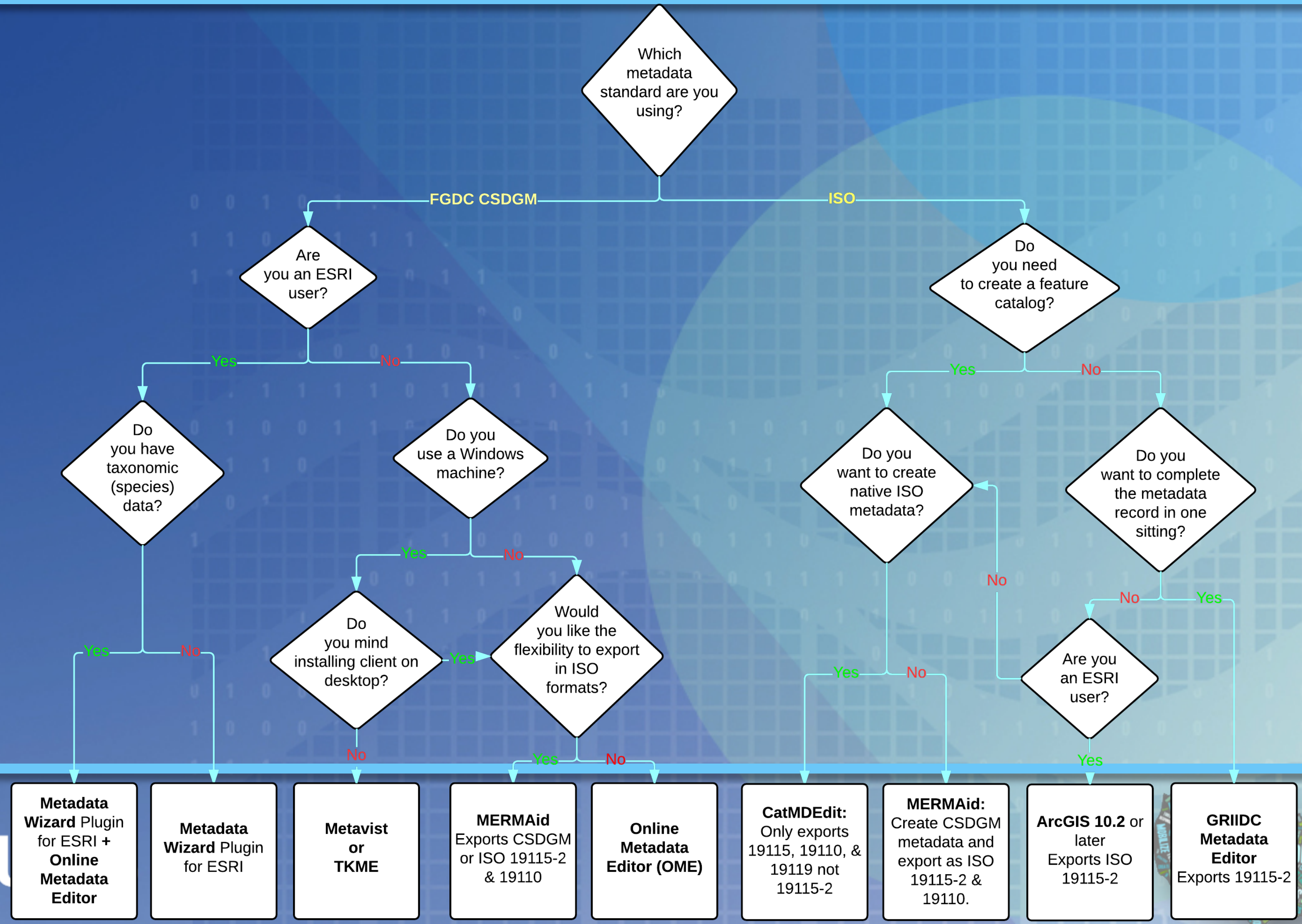
CSDGM tools

COMMERCIAL:

✓ ArcGIS Desktop %
✓ SMMS (desktop) #

# Supports Biological Data Profile
^ Supports Biological Data Profile, Shoreline Profile, and Remote Sensing Extension
% Provides validation

Metadata Tool Guidance



Which metadata standard are you using?

Are you an ESRI user?

Do you have taxonomic (species) data?

Do you use a Windows machine?

Do you mind installing client on desktop?

Would you like the flexibility to export in ISO formats?

Do you need to create a feature catalog?

Do you want to create native ISO metadata?

Do you want to complete the metadata record in one sitting?

Are you an ESRI user?

Metadata Wizard Plugin for ESRI **+ Online Metadata Editor**

Metadata Wizard Plugin for ESRI

Metavist or TKME

MERMAid Exports CSDGM or ISO 19115-2 & 19110

Online Metadata Editor (OME)

CatMDEdit: Only exports 19115, 19110, & 19119 not 19115-2

MERMAid: Create CSDGM metadata and export as ISO 19115-2 & 19110.

ArcGIS 10.2 or later Exports ISO 19115-2

GRIIDC Metadata Editor Exports 19115-2

Best practice: start early

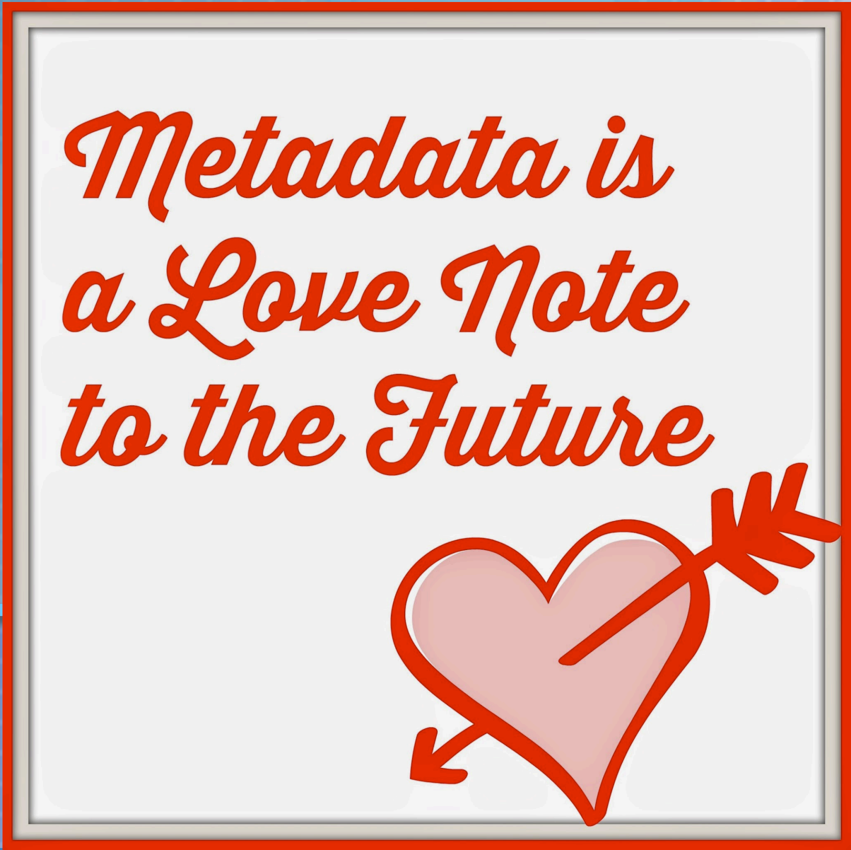
- ✓ Don’t put off metadata until the end of your project
- ✓ Documenting as you go ensures that details aren’t lost
- ✓ Yes, data change during the project
  - ✓ Editing metadata later is easier than starting from scratch at the end

Best practice: fully document your data

✓ Completing only the minimum required elements virtually ensures that your data won’t be reusable later

✓ Full metadata documents the integrity of your research

✓ Who can predict what data will be valuable 10, 20, or 50 years from now?



Best practice: adopt consistent practices across your program or agency

- ✓ Syntax and structure rules
  - ✓ Specifying organizational names
    - ✓ E.g., USGS vs. U.S.G.S. vs. U.S. Geological Survey vs. United States Geological Survey
  - ✓ Syntax for personal names
    - ✓ Doe, John vs. Doe, J. vs. John Doe
  - ✓ Use of acronyms and highly specific jargon

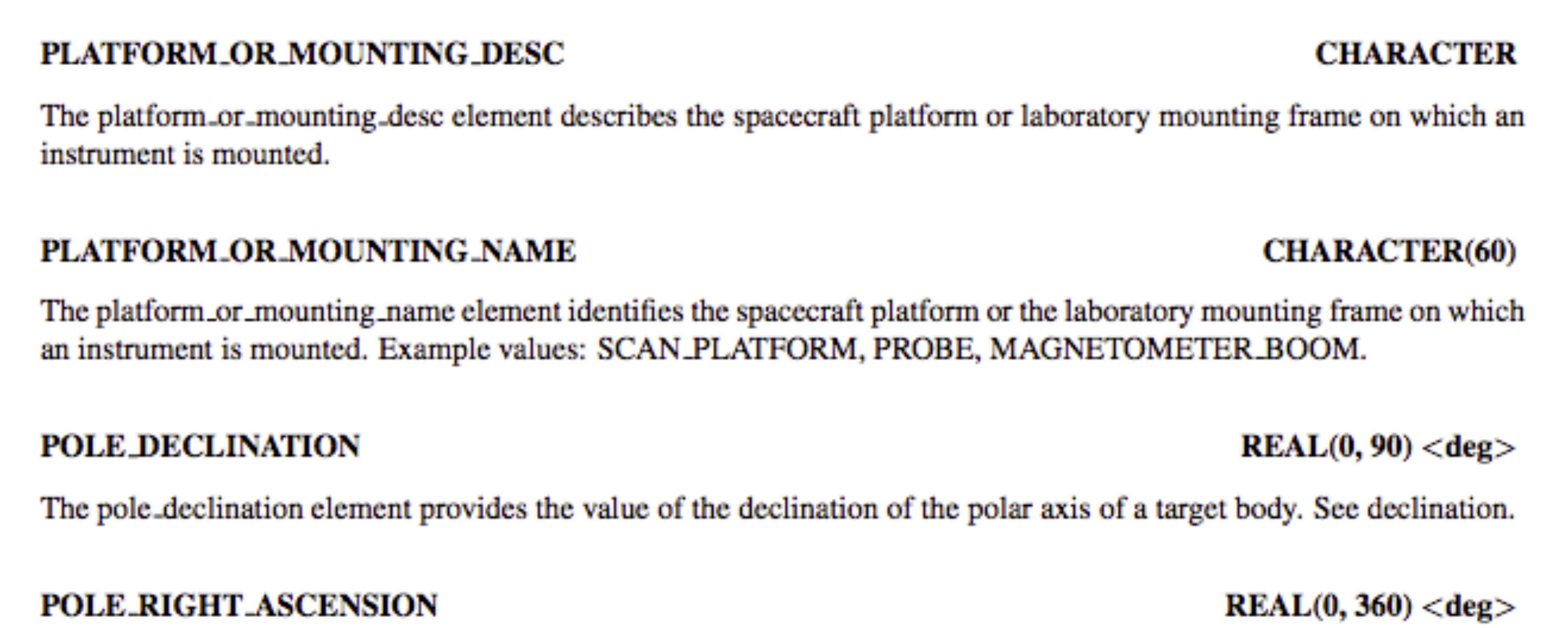
Best practice: adopt consistent practices across your program or agency

- ✓ Syntax and structure rules
  - ✓ Use of a common, authoritative terminology systems
    - ✓ Subject keywords, place names, species names
    - ✓ Federal science agencies' terminology systems (From CENDI)
  - ✓ Repeatable elements need to be placed in individual XML tags
    - ✓ NO: <origin>John Doe, Jane Smith, Jean Johnson</origin>
    - ✓ YES: <origin>John Doe</origin>
      <origin>Jane Smith</origin>
      <origin>Jean Johnson</origin>

USGS

Best practice: consider creating a data dictionary

- ✓ Entity-Attribute/Feature Catalog metadata =labor-intensive
- ✓ If your datasets are relatively homogeneous, developing a data dictionary can save a lot of time
- ✓ Reference the data dictionary in lieu of detailed E/A



PLATFORM_OR_MOUNTING_DESC CHARACTER

The platform_or_mounting_desc element describes the spacecraft platform or laboratory mounting frame on which an instrument is mounted.

PLATFORM_OR_MOUNTING_NAME CHARACTER(60)

The platform_or_mounting_name element identifies the spacecraft platform or the laboratory mounting frame on which an instrument is mounted. Example values: SCAN_PLATFORM, PROBE, MAGNETOMETER_BOOM.

POLE_DECLINATION REAL(0, 90) <deg>

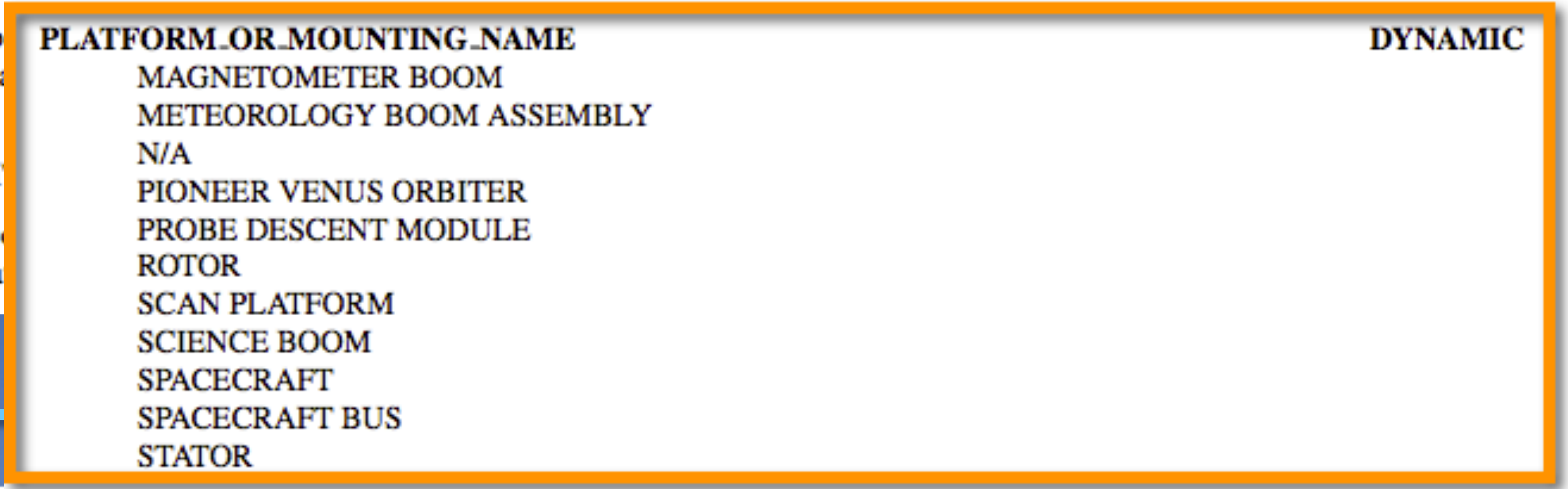
The pole_declination element provides the value of the declination of the polar axis of a target body. See declination.

POLE_RIGHT_ASCENSION REAL(0, 360) <deg>

The p
right_a

POSI

The p
forma



PLATFORM_OR_MOUNTING_NAME DYNAMIC

MAGNETOMETER BOOM
METEOROLOGY BOOM ASSEMBLY
N/A
PIONEER VENUS ORBITER
PROBE DESCENT MODULE
ROTOR
SCAN PLATFORM
SCIENCE BOOM
SPACECRAFT
SPACECRAFT BUS
STATOR

Example: Planetary Science Data Dictionary Document (NASA)

Best practice: don’t skip Data Quality

- ✓ Account for accuracy and completeness of your data
- ✓ Lineage is extremely important!
  - ✓ Methods (for biological data)
  - ✓ Source Information (other datasets used and their provenance)
  - ✓ Processing steps for the dataset (raw → finished dataset)

Best practice: describe the spatial resolution of your data appropriately

- ✓ Specify units
- ✓ For a range of locations in your data
  - ✓ Overall bounding coordinates AND
  - ✓ Describe the data acquisition scheme textually (Description of Geospatial Extent)
    - ✓ E.g., "Data were collected every 100m along a series of 500m transects (defined in associated shapefile)."

Best practice: document all units of measurement

- ✓ Don’t leave users guessing about any measurements
- ✓ Document these thoroughly in Entity-Attribute/Feature Catalog
- ✓ Cautionary Tale: Mars Orbiter

Best practice: assign a persistent identifier to your data and manage it

- ✓ The online location of final data is likely to change at some point
  - ✓ Usually metadata and other finders have to be updated in many places
  - ✓ Obtain a DOI and use the resolvable DOI in place of a URL for your data in the metadata
  - ✓ Manage the data’s location in DOI registry
  - ✓ Change once, propagate many
- ✓ DOIs support appropriate citation of your data
- ✓ DOIs support citation analysis, ROI of data beyond their *original* purpose and use

USGS

Best practice: complete Access Constraints, Use Constraints, and Distribution Liability

- ✓ Access Constraints
  - ✓ For federal datasets, open access is assumed. Any other condition must be explained.
- ✓ Use constraints
  - ✓ Particularly important if you're using 3rd party or licensed data
- ✓ Distribution Liability
  - ✓ Determine whether your agency has preferred language to protect from unintended uses of, or assumptions about, the data.

USGS

Best practice: validate your metadata!

- ✓ Many metadata tools provide no validation
- ✓ A huge percentage of existing metadata records contain errors in structure, syntax, and content
  - ✓ Risks: catalog harvest failures, errors in understanding the data, conversion failures between CSDGM and ISO
- ✓ If your metadata tool provides no validation:
  - ✓ For CSDGM: use MP
  - ✓ For ISO: use Schematron

Best practice: have your data and metadata reviewed together

- ✓ Have a knowledgeable 3rd party review your data *and* metadata prior to release
- ✓ This is your best chance for catching omissions and errors that can inhibit both discovery and understanding of your data
  - ✓ Discovery examples: typos; bad bounding coordinates
  - ✓ Understanding examples: Missing process step; bad or undefined attribute

Best practice: package your metadata with your data

- ✓ Deposit all metadata and documentation with your data at the delivery point
- ✓ Zipping metadata with data ensures it will always be downloaded with the data

USGS

Best practice: publish your metadata

- ✓ Data aren’t discoverable if metadata aren’t shared
- ✓ Publish your XML metadata file to the appropriate program, agency, or discipline catalogs
- ✓ Find out how your metadata can be published to data.gov and geoplatform.gov

Best practice: think of metadata as living documents

- ✓ Data and metadata must be maintained for the useful life of the data
- ✓ Update metadata as necessary to reflect changes to the data (updates, corrections, status); points of contact; distribution; etc.

Thoughts? Questions?

Thanks for your interest in metadata today!

lisa_zolly@usgs.gov

USGS